DOCUMENT RESUME

ED 218 306 TM 820 340

AUTHOR Holmes, Susan E.
TITLE The Effects of Test Content Match and Number of Items on the Accuracy of Trait Estimates from Tests Equated with the Three-Parameter Logistic Model.
PUB DATE 21 Mar 82
NOTE 36p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Content Analysis; Elementary Education; *Equated Scores; Item Analysis; *Latent Trait Theory; Mathematical Models; Test Theory; *Test Validity
IDENTIFIERS *California Achievement Tests; *Prescriptive Reading Inventory; Three Parameter Model

ABSTRACT

The purpose of the present study was to examine the accuracy of indirect trait estimates, i.e., estimates of some primary trait obtained from a second measure which have been equated to the first. The California Achievement Test in Reading was the primary measure and the Prescriptive Reading Inventory was the indirect measure. Four kinds of content-match between primary and indirect item sets were studied: match, similar, dissimilar, and partial. Within each content-match category, indirect trait estimates based on item sets containing 10, 20, and 30 items were compared to those obtained from the primary measure. A three-parameter logistic model was applied to an internal anchor test design to equate the measures. The results obtained were then evaluated for a cross-validation sample. The accuracy of indirect trait estimates was affected by both content-match and the number of items used to produce the estimate. As indirect item sets became more similar in content or as the number of items increased, the accuracy of the estimates increased. However, a considerable amount of equating bias was noted, even for the 30-item sets. (Author/CM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED218306

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

The Effects of Test Content Match and Number of Items on the Accuracy of Trait Estimates from Tests Equated with the Three-Parameter Logistic Model

Susan E. Holmes
CTB/McGraw-Hill

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

S. E. Holmes

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 820 310

A paper presented at the Annual Meeting of the American Educational Research Association, March 21, 1982

Introduction

Contemporary psychometric literature has been dominated by research on item response theory (IRT) models. This has been the case because of the potential of these models for solving a variety of intractable measurement problems (Hambleton, Swaminathan, Cook, Eignor & Gifford, 1978; Marco, 1977). Test equating is one of the areas in which the application of IRT models has been most promising (Lord, 1975,1977).

Test equating is the process of converting the system of units of one test to the system of units of a second so that scores obtained from the two tests are equivalent after conversion (Angoff, 1971). Two tests may then be said to be equated if it no longer matters to those being tested which of the two they actually take (Lord, 1980). In other words, two tests may be considered to be successfully equated when performance is not affected by the particular test form being administered. Interchangeable forms of the same test are often a requirement in large scale testing programs. Test equating is necessary because conventional test construction techniques can not guarantee that two tests will be equivalent in level and/or range of difficulty.

One of the requirements of the test equating process is that the tests being equated measure the same trait or characteristic. This assumption seems reasonable for published test batteries which include by design parallel forms of the same test. In many practical situations, however,

test equating is often performed on tests which were not specifically designed to be parallel. For example, two reading tests developed by different test publishers or a reading test developed by a test publisher and one developed by a local school district may be equated. In many instances the tests being equated differ with respect to the content measured as well as the number of items used to measure that content. In most cases, however, the intent of the test equating is the same: to obtain an estimate of performance on one test (primary measure) by examining performance on a second test (indirect measure) which has been equated to the first.

The use of IRT models for test equating offers several advantages over more conventional test equating techniques. First, this approach does not require that the same or equivalent groups of examinees take the tests being equated. This is the case because of the invariance property of IRT model item parameters.

A second advantage of using IRT models for test equating is that after equating, all items on the tests being equated are calibrated on a common trait scale. Theoretically, then, any subset of them could subsequently be used to estimate the trait of interest. An appealing combination of items from the perspective of the local school district might be some subset selected from a locally developed test, particularly if the test to which it had been equated was a published achievement test. Once the equating had

been completed, an estimate of performance on the published test, could be obtained from any administration of the locally developed test simply by using the equated item parameters for those items selected. If administrations were conducted at appropriate times during the school year, the national normative information associated with the published test might also be associated with the item set selected. This additional information would be obtained at a considerable savings in administration time and cost to the local district.

While a testing strategy such as the one outlined above is possible using IRT models, the concern about equating two tests which may not measure the same trait still exists. The accuracy of "indirect" trait estimates obtained in this way would be affected by the content-match between the two tests that were initially equated as well as between the primary measure and the particular item subset selected from the indirect measure. Systematic content selection in identifying the item subset could result in biased indirect trait estimates. The number of items included in the subset would also be an important factor. Little research about the implications of such a testing strategy has appeared in the literature.

The purpose of the present study was to examine the accuracy of indirect trait estimates, i.e. estimates of some primary trait obtained from a second measure which have been equated to the first. For the purposes of this paper, the trait measured by the first item set will be referred to as the primary trait estimate. An indirect trait estimate is therefore an estimate of a primary trait obtained (indirectly) from a second measure. Of particular interest was the effect on accuracy of the degree to which the primary

and indirect item sets actually measure the same thing. Four kinds of content-match between primary and indirect item sets were therefore studied: match, similar, dissimilar and partial. Within each of these content-match categories, indirect trait estimates based on item sets containing 10, 20 and 30 items were compared to those obtained from the primary measure. A three-parameter logistic (IRT) model was applied to an internal anchor test design to accomplish the test equating. The results obtained were then evaluated for a cross-validation sample.

Method

Data

Two data sets were studied, one at each of two grade levels (Grades 2 and 6). Each data set consisted of the item responses from a pair of reading tests. The first member in the pair was the California Achievement Tests (1977), Form C (CAT). CAT is a battery of achievement tests that measures the basic curricular areas of reading, language and mathematics. Data from only the reading test were included in this study. The second member in the test pair was the Prescriptive Reading Inventory (1972) (PRI). PRI is a criterion-referenced test battery that measures mastery of commonly taught reading objectives.

The data studied were originally collected as part of the 1977 standarization of CAT. Table 1 presents the number of reading items in each test pair as well as the sample size available from the standardization. For the purposes of this paper, CAT will be referred to as the primary measure and PRI as the indirect measure.

Insert Table 1 about here

Two random samples were selected from each data set. The first sample, referred to as the "equating sample", was used to perform the test equating. The results obtained from the equating were then applied to the second sample, referred to as the "cross-validation" sample. All samples contained 996 cases except the cross-validation sample at Grade 2. This sample contained only 905 cases.

Selection of Anchor Items

In order to examine the accuracy of indirect trait estimates, various PRI item sets were identified and then equated to the CAT scale via an internal anchor design. Trait estimates from these item sets (indirect trait estimates) were then compared to those obtained from CAT (primary trait estimate.) What follows is a description of the anchor item selection process.

An anchor item set consisting of 20 items was identified for both of the primary measures described in Table 1. Initially, each primary measure was analyzed using LOGIST (Wood, Wingersky, & Lord, 1976). Item discriminating power (a_i), item difficulty (b_i) and the lower asymptote, or pseudo-chance-success level, (c_i) for the three-parameter logistic model were estimated for each item (i). Only the equating sample data were used.

Using the normative data available for the primary measures (California Achievement Tests: Norms Tables, Level 12 and 16, 1977), a range of item difficulty within which an adequate degree of measurement accuracy might be expected was established for each measure. These values were based on performance at the 25th and 75th percentiles, respectively, at the lowest and highest grade for which each measure was appropriate. Anchor items were then selected proportionately by objective from among those items within the difficulty range for that measure. For example, the level 12 CAT reading test includes both vocabulary and comprehension items. Ten vocabulary and 10 comprehension items were therefore selected. Since the vocabulary items measured two objectives (Same Meaning, Opposite Meaning), five anchor items were selected from among those measuring each of these two objectives.

Where possible, items were selected from across the entire difficulty range. The intention of this procedure was to identify internal anchor item sets which reflected as closely as possible the content and difficulty of the primary measures from which they were selected.

Selection of Indirect Item Sets

Item sets containing 10, 20 or 30 items with a variety of content-matches to CAT were identified from among those in PRI. These item sets were the indirect measures used to estimate the appropriate primary (CAT reading) trait after they had been equated to the CAT scale. What follows is a description of how these item sets were selected.

Four content-match categories were considered: match, similar, dissimilar and partial. PRI items were categorized according to the mapping of CAT to PRI objectives presented in the CAT Class Management Guide (1978, pp. 163-169). The following definitions were applied:

Match - PRI item sets that measured objectives identical to those measured in CAT. All objectives included in CAT were represented within each item set. These item sets were the "best" available matches between primary and indirect measures.

Similar - PRI item sets that measured objectives similar but not identical to those measured in CAT. (An example might be items which measure PRI objectives involving words such as Inflected Words: Singular/Plural or Pronouns. These objectives might be considered similar to a CAT objective like Same Meaning which also involves words.)

Dissimilar - PRI item sets that measured objectives not measured in CAT. (An example would be PRI items measuring the objective Vowel Sounds: Long and Short which is not included in CAT).

Partial - PRI item sets that measured objectives identical to those measured in CAT, but which measured only half of all those objectives included in CAT.

Twelve item sets were identified at each grade, a 10, 20 and 30 item set in each of the four content-match categories above. Initially, all PRI items were assigned to one of the first three content-match categories. From those items available in each category, a 30 item set was then selected. The match and similar item sets were selected to reflect proportionately all the objectives included in CAT. The partial item sets were selected from among those items in the match category. They contained items measuring only the comprehension objective included in CAT. The intention was to

create an indirect item set which matched CAT very well, but only for some objectives. A 10 and 20 item set was then identified from among the 30 selected in each content-match category.

Evaluation of Content-Match

In order to examine the content-match between CAT and PRI, an independent item by item rating of the two tests was performed by a reading specialist. The review process consisted of comparing each PRI item with every CAT item in a particular test pair and then rating the match between the content measured by the two items. The following scale was used:

5 = same content measured in the same way (same format)

4 = same content measured but in a different way

3 = similar content measured in the same way

2 = similar content measured but in a different way

1 = different content measured in the same way

0 = different content measured in a different way

The percent of items receiving each rating was then determined for each of the 12 indirect item sets at each grade. This information was used to evaluate the degree to which the content measured by the indirect item sets selected actually matched the content measured by the primary measure.

Test Equating

In order to compare trait estimates obtained from primary and indirect item sets, all pairs of measures were equated using a three-parameter model equating technique. Item parameters were estimated for the 12 indirect item sets and the CAT anchor test in a single analysis using LOGIST and the equating sample data. The linear relationship between the anchor item difficulty estimates (b_i's) obtained from this analysis and those obtained from an analysis of only the CAT items was then used to place the item parameter estimates for the 12 indirect item sets on the CAT scale (Warm, 1978, p. 113-114).

Primary and Indirect Trait Estimates

One primary and twelve indirect trait estimates were then generated for each examinee in both samples. The appropriate combination of re-scaled item parameter estimates was applied to the item response data of each examinee in order to produce each indirect trait estimate. These estimates were then compared to the estimate obtained from applying only the CAT item parameter estimates. It was possible to generate pairs of estimates for each examinee because all examinees in both samples had actually responded to all PRI and CAT items.

The effect of test content-match and number of items on the accuracy of indirect trait estimates was then evaluated by examining the Pearson product-moment correlation between pairs of primary and indirect trait estimates.

Further, a root mean squared difference (RMSD) between pairs of trait estimates was computed. This value, scaled by the variances of the two trait estimates being compared, is given as follows:

$$\frac{RMSD}{S_{\hat{\theta}}} = \frac{\sqrt{\frac{1}{n}\sum_{j=1}^{n}(\hat{\theta}_{1j} - \hat{\theta}_{2j})^2}}{\sqrt{(S^2_{\hat{\theta}_1} + S^2_{\hat{\theta}_2})/2}} .$$

Here $\hat{\theta}_{1j}$ and $\hat{\theta}_{2j}$ are two estimates of the same trait for person "j" (j = 1,n examinees), and $S^2_{\hat{\theta}_1}$ and $S^2_{\hat{\theta}_2}$ are the variances of the two estimates. RMSD values can be interpreted as empirical standard errors and were rescaled to facilitate comparisons across item sets.

Results

Content-Match Definitions

In order to evaluate the actual content-match between the various indirect item sets and the appropriate primary measure, the content measured by each PRI item was compared to the content measured by every CAT item and then rated according to match. The results of this review process are presented in Table 2, where the percent of ratings are summarized in three categories. Same, Similar and Different. Since each PRI item was compared to every item on the primary measure, the values reported in Table 1 represent the percent of all these ratings in each category.

Insert Table 2 about here

Summary Test Statistics

Summary statistics for the two primary measures studied are reported by grade in Table 3 for both the equating and cross-validation samples. Included in this table are the raw score mean and standard deviation, as well as the Pearson product-moment correlation between each primary measure and the 20 item anchor test identified for the equating.

Insert Table 3 about here

Summary statistics for the 12 indirect item sets studied at each grade are presented in Tables 4 and 5 for the equating and cross-validation samples, respectively. Raw score means and standard deviations appear there, as well as the Pearson product-moment correlations between pairs of indirect and primary measure raw scores. The average proportion of correct responses ($\overline{p}$) included in Table 4 is based on item values reported for the CAT standardization sample (California Achievement Tests: Technical Bulletin 1, 1979). Average proportion correct values were computed for the 10, 20, and 30 item sets in each content-match category to determine if these sets differed with respect to difficulty. The values reported in Table 4 indicate that while the average proportion of correct responses increased as the number of items in the indirect item set increased, the size of the increase was small.

Insert Table 4 and 5 about here

Primary Trait Estimates

Table 6 presents summary statistics for the trait estimates obtained from each of the primary measures studied and from the anchor item sets. The means and standard deviations are reported by grade and sample. The values presented for the cross-validation sample were obtained by applying the item parameter estimates generated from the equating sample data to the item response data from the cross-validation sample. Also reported in Table 6 is the Pearson product-moment correlation between pairs of trait estimates, one estimate being obtained from the primary measure and the other from its anchor. Examinees with zero or perfect scores on either item set in a

particular pair were excluded from the computations reported in Table 6. Trait estimates can not be generated for examinees obtaining these scores using maximum likelihood estimation procedures.

Insert Table 6 about here

Indirect Trait Estimates

Since all examinees had actually responded to all items in both the primary and indirect measures, pairs of trait estimates could be generated for each examinee in both samples after the two measures had been equated. The primary and indirect trait estimate means and standard deviations obtained are reported by grade in Tables 7 and 8 for the equating and cross-validation samples, respectively. In addition, the correlation and (scaled) root mean square difference (RMSD) for pairs of trait estimates are reported there. The latter can be viewed as an empirical standard error for the difference between pairs of trait estimates.

Insert Tables 7 and 8 about here

In order to facilitate comparisons, the correlations reported in Tables 7 and 8 are plotted in Figures 1 and 2, respectively. Again, examinees with zero or perfect scores on either measure in a pair were excluded. Examinees whose trait estimates had extremely large standard errors were felt to be poorly measured. For this reason examinees whose trait estimates had standard errors more than five times greater than the minimum value obtained for either member in the test pair were excluded.

Insert Figures 1 and 2 about here

Discussion

The purpose of the study was to examine the accuracy of indirect trait estimates, i.e. estimates of some primary trait obtained from a second measure which had been equated to the first. Two variables were examined: the content-match between the primary and indirect measure and the number of items used to obtain the indirect trait estimate.

The intention of the study was to examine the feasibility of a particular testing strategy made possible by the use of IRT. The strategy of interest was one in which a subset of items from a locally developed test (indirect measure) is used to estimate performance on another test (primary measure), perhaps a nationally normed publisher's test proported to measure the same thing Initially, the two tests were equated using a three-parameter logistic model. Trait estimates based on various subsets of items from the indirect measure were then compared to an estimate based on the primary measure. In this way the effect on accuracy of systematic content selection when identifying the subset of indirect items was examined. The effect of the number of items used to generate the indirect estimate was also examined.

Content-Match Definitions

Indirect item sets were selected to have one of four kinds of content-match to the primary measure. The percents reported in Table 2 suggest that this was generally the case. The "match" item sets had the largest percent of "same" ratings (4 or 5), with the similar and partial item sets sharing the next largest percents and the dissimilar item sets having the smallest. Conversely, the dissimilar item sets had the largest percents of "different"

ratings (0 or 1). Further, the average proportions correct (Table 4) suggest that while some differences in difficulty existed between and within content-match categories, they were not substantial.

It should be noted, however, that differences between content-match categories (Table 2) were not great. This is not surprising since all indirect item sets contained reading items, even though those items may have measured reading objectives not included in the primary measure. The raw score correlations reported for the equating sample (Table 4), particularly those at Grade 6, also support this observation. Further, as the number of items in the indirect item set increased, the percent of items with a "same" or "similar" rating decreased. This suggests that a consistent definition of content-match was not maintained within some content-match categories, particularly those at Grade 2.

Test Equating

Twenty anchor items were selected from each primary measure in order to perform the equating. The correlations between anchor item sets and primary measures (Table 3) suggest that the anchor item sets did reflect the primary measures from which they were chosen, although the correlations reported are inflated. This conclusion is also supported by the mean anchor and primary trait estimates reported in Table 6. The correlations between estimates presented there are also inflated.

The results of the test equating process itself warrant some discussion. An examination of the pairs of mean trait estimates reported in Table 7 suggests that some bias was present in the test equating. At Grade 2, consistently

lower average values were obtained from indirect measures than from the primary measure. The reverse was true for Grade 6. Further, these differences became more pronounced as the number of items in the indirect item set decreased.

The phenomenon noted above is probably due to the number of examinees at each grade level excluded because of zero or perfect scores on the indirect item set in the pair being examined. In general, the indirect item sets at Grade 2 were easier than those at Grade 6 (see the average proportions correct in Table 4). This resulted in a larger proportion of examinees being deleted because of perfect scores at this grade level than at Grade 6. In general, fewer examinees were deleted at Grade 6 than at Grade 2 and more of them were deleted because of zero rather than perfect scores. As the number of items in the indirect item set decreased, this difference became increasingly more pronounced, resulting in the pattern of average trait estimate pairs in Table 7. It is hypothesized that this kind of bias would disappear if floor and ceiling trait estimate values were established. A re-analysis of the data implementing such an approach is planned.

Content-Match and Accuracy

In general, the results reported for the equating sample in Table 7 and Figure 1 indicate that the accuracy of indirect trait estimates is affected by the content-match between the subset of items used to estimate the trait (the indirect item set) and the primary measure. Indirect item sets labeled "match" generally produced the highest correlations (and therefore the lowest empirical standard errors) between primary and indirect trait estimates.

"Partial" and "similar" item sets produced the next highest correlations, with "dissimilar" item sets producing the lowest correlations. This was more evident at Grade 6 than at Grade 2, where consistently higher correlations were reported. It should be noted that the differences among raw score correlations reported in Table 6 were not great. This suggests that the differences in content-match implied by the item set labels may not have been large. Larger differences might have been obtained if more distinct content-match categories had been studied or if the indirect item sets had been individually equated to the primary measure.

Number of Items and Accuracy

As would be expected, the accuracy of indirect trait estimates was effected by the number of items used in computing those estimates. As can be seen in Table 7 and Figure 1, 30 items consistently produced the highest correlations between primary and indirect trait estimates, with 20 items producing the next highest and 10 items producing the lowest. This was the case regardless of grade level or content-match between the two measures. The largest increase, however, was between 10 and 20 items. These data suggest that the increased testing time required by increasing the number of items from 20 to 30 might be unwarranted given the small increase in accuracy the additional items would provide.

Cross-Validation

The accuracy of indirect trait estimates was also examined using cross-validation samples. These samples were randomly selected from among those cases available in each data set studied (see Table 1). The raw score means,

standard deviations and correlations reported in Table 3 and in Tables 4 and 5 indicate that the equating and cross-validation samples were in fact quite similar.

After equating, the item parameter estimates obtained from the equating sample data were applied to the item response data for the cross-validation samples in order to produce pairs of primary and indirect trait estimates. The values obtained (Table 8) are consistent with those reported for the equating sample (Table 7) at both grade levels. However, the cross-validation correlations are slightly lower, particularly at Grade 6. Larger differences might be expected from a non-random cross-validation sample.

Summary

In general, the accuracy of indirect trait estimates was affected by both content-match and the number of items used to produce the estimate. As indirect item sets became more similar in content to the primary measure for which they were an estimate, accuracy tended to increase. Accuracy also increased as the number of items used in generating the estimate increased, with the largest increase occurring between 10 and 20 items. However, a considerable amount of equating bias was noted, even for the 30 item sets. Results consistent with these conclusions were also found for the cross-validation samples.

References

Angoff, W. H. Scales, norms, and equivalent scores. In Thorndike (Ed.), Educational measurement (2nd ed.). Washington, DC: American Council on Education, 1971.

California achievement tests: Norms tables, level 12. Monterey, CA: CTB/McGraw-Hill, 1977.

California achievement tests: Norms tables, level 16. Monterey, CA: CTB/McGraw-Hill, 1977.

California achievement tests. Monterey, CA: CTB/McGraw-Hill, 1977.

California achievement tests: Technical bulletin 1. Monterey, CA: CTB/McGraw-Hill, 1979.

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J. A. Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research. 1978, 48, 467-510.

Lord, F. M. A survey of equating methods based on item characteristic curve theory (ETS RB 75-13). Princeton, NJ: Educational Testing Service, 1975.

Lord, F. M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates, 1980.

Marco, G. L. Item characteristic curve solutions to three intractable testing problems. Journal of Educational Measurement, 1977, 14, 139-160.

Warm, T. A. A primer of item response theory (CG-941278). U.S. Coast Guard Institute: P.O. Substation 18, Oklahoma City, Oklahoma, December 1978.

Wood, R. L., Wingersky, M. S., & Lord, F. M. A computer program for estimating examinee ability and item characteristic curve parameters (ETS RM 76-6). Princeton, NJ: Educational Testing Service, 1976.

Table 1

Two Test Pairs Studied, the Number of Items in Each, the Grade for which Each Is Appropriate, and the Number of Available Cases

Primary Measure		Indirect Measure			
Description	# Items	Description	# Items	Grade	N
CAT 12 Reading Test	35	PRI-A	123	2	1947
CAT 16 Reading Test	70	PRI-D	148	6	3963

Table 2

Percent of Items in Each Indirect Item Set
by Rating Category and by Grade

Content-Match Category		Grade 2			Grade 6		
		4 or 5 (Same)	2 or 3 (Similar)	0 or 1 (Different)	4 or 5 (Same)	2 or 3 (Similar)	0 or 1 (Different)
Match	10	11.1	34.6	54.3	11.0	27.6	61.4
	20	9.1	32.3	58.6	10.9	28.4	61.4
	30	7.0	30.1	62.9	8.2	31.6	60.2
Similar	10	2.0	26.6	71.4	3.1	34.0	62.9
	20	1.6	24.1	74.3	3.3	31.0	65.7
	30	1.1	19.8	79.1	3.4	33.7	62.9
Dissimilar	10	0.0	0.0	100.0	.6	28.0	71.4
	20	0.0	0.0	100.0	.7	27.9	71.4
	30	0.0	0.0	100.0	.6	28.0	71.4
Partial	10	5.1	52.0	42.9	1.5	55.6	42.9
	20	4.5	46.9	48.6	1.6	55.5	42.9
	30	4.8	46.6	48.6	1.7	55.4	42.9

Table 3

Primary and Anchor Test Raw Score Means, Standard Deviations and Correlations by Grade and Sample

		Primary Measure			Anchor Test[a]		
Grade	N	# Items	$\bar{X}_1$	S_{X1}	$\bar{X}_2$	S_{X2}	$r_{X_1X_2}$
Equating Sample							
2	996	35	20.59	9.60	12.09	5.68	.98
6	996	70	42.84	15.19	12.18	4.84	.95
Cross-Validation Sample							
2	905	35	20.96	9.51	12.52	5.51	.98
6	996	70	43.08	14.58	12.19	4.68	.95

[a]All anchor tests contained 20 items.

Table 4

Indirect Item Set Raw Score Means, Standard Deviations, Proportions Correct and Correlations with Primary Measure by Grade (Equating Sample)[a]

Content-Match Category	Indirect Item Set # Items	$\bar{X}$	S_X	$\bar{p}$[b]	$\underline{r}$
		Grade 2			
Match	10	7.25	2.59	.65	.68
	20	14.41	4.69	.71	.71
	30	21.86	6.77	.73	.73
Similar	10	7.35	2.62	.66	.70
	20	14.29	4.66	.70	.74
	30	20.89	7.10	.70	.76
Dissimilar	10	7.61	2.05	.73	.61
	20	14.86	3.81	.74	.65
	30	22.51	5.65	.75	.67
Partial	10	7.10	2.44	.67	.63
	20	14.34	4.53	.71	.69
	30	21.54	6.63	.73	.71
		Grade 6			
Match	10	5.88	2.60	.58	.76
	20	11.67	4.72	.59	.80
	30	17.11	6.77	.58	.82
Similar	10	5.68	2.13	.58	.67
	20	11.34	3.69	.58	.74
	30	16.68	5.56	.56	.78
Dissimilar	10	5.72	2.14	.57	.68
	20	11.95	4.22	.61	.76
	30	17.54	6.12	.60	.79
Partial	10	5.35	2.56	.53	.73
	20	10.42	4.63	.53	.79
	30	15.97	6.86	.54	.81

[a] N = 996.

[b] California Achievement Tests: Technical Bulletin 1, 1979.

Table 5

Indirect Item Set Raw Score Means, Standard Deviations and Correlations with Primary Measure by Grade (Cross-Validation Sample)[a]

Content-Match Category	Indirect Item Set # Items	$\bar{X}$	S_X	r
Grade 2				
Match	10	7.30	2.62	.69
	20	14.63	4.85	.73
	30	22.16	6.89	.75
Similar	10	7.43	2.58	.71
	20	14.61	4.54	.74
	30	21.32	6.96	.77
Dissimilar	10	7.72	2.10	.63
	20	14.99	3.88	.66
	30	22.63	5.82	.69
Partial	10	7.17	2.47	.66
	20	14.52	4.64	.70
	30	21.94	6.65	.72
Grade 6				
Match	10	6.00	2.45	.73
	20	11.84	4.49	.77
	30	17.35	6.52	.80
Similar	10	5.77	2.04	.68
	20	11.49	3.60	.74
	30	16.87	5.46	.78
Dissimilar	10	5.90	2.08	.66
	20	12.23	4.11	.78
	30	17.95	6.02	.80
Partial	10	5.41	2.45	.70
	20	10.51	4.55	.76
	30	16.10	6.71	.79

[a] N = 905 for Grade 2 and N = 996 for Grade 6.

Table 6

Primary and Anchor Trait Estimate Means, Standard Deviations, and Correlations by Grade and Sample

			Primary Trait Estimate		Anchor Trait Estimate[a]		
Grade	N[b]	# Items	$\bar{\hat{\theta}}_1$	$S_{\hat{\theta}_1}$	$\bar{\hat{\theta}}_2$	$S_{\hat{\theta}_2}$	$r_{\hat{\theta}_1\hat{\theta}_2}$
			Equating Sample				
2	836	35	-.02	.84	-.03	.84	.97
6	865	70	.08	.85	.10	.88	.91
			Cross-Validation Sample				
2	777	35	.06	.85	.05	.83	.97
6	886	70	.06	.81	.06	.88	.91

[a]All anchor item sets contained 20 items.

[b]Cases with zero or perfect scores on either item set in a pair were excluded from these computations.

Table 7

Primary and Indirect Trait Estimate Means, Standard Deviations, Correlations, and Root Mean Squares by Grade (Equating Sample)[a]

Content-Match Category	N	Primary Trait Estimate $\bar{\hat{\theta}}_1$	Primary Trait Estimate $S_{\hat{\theta}1}$	# Items	Indirect Trait Estimate $\bar{\hat{\theta}}_2$	Indirect Trait Estimate $S_{\hat{\theta}2}$	$r_{\hat{\theta}_1\hat{\theta}_2}$	$\frac{RMSD}{S_{\hat{\theta}}}$
				Grade 2				
Match	629	-.17	.84	10	-.47	.83	.64	.92
	781	-.01	.86	20	-.16	1.02	.70	.79
	829	-.00	.88	30	-.15	1.06	.75	.74
Similar	609	-.19	.82	10	-.48	.76	.63	.94
	781	-.01	.86	20	-.19	.94	.73	.77
	823	.03	.88	30	-.13	.97	.78	.75
Dissimilar	593	-.24	.83	10	-.60	.86	.58	1.00
	776	-.05	.89	20	-.20	1.18	.69	.82
	801	-.04	.88	30	-.21	1.11	.71	.81
Partial	708	-.04	.87	10	-.23	1.02	.61	.91
	815	-.01	.89	20	-.16	1.14	.71	.78
	845	.02	.90	30	-.12	1.10	.76	.73

Table 7
(continued)

Content-Match Category	N	Primary Trait Estimate $\bar{\hat{\theta}}_1$	Primary Trait Estimate $S_{\hat{\theta}1}$	# Items	Indirect Trait Estimate $\bar{\hat{\theta}}_2$	Indirect Trait Estimate $S_{\hat{\theta}2}$	$r_{\hat{\theta}_1\hat{\theta}_2}$	$\frac{RMSD}{S_{\hat{\theta}}}$
				Grade 6				
Match	799	.08	.83	10	.16	.80	.70	.78
	862	.11	.88	20	.20	.88	.78	.66
	866	.13	.88	30	.24	.86	.82	.62
Similar	889	.06	.94	10	.23	1.14	.65	.87
	915	.06	.95	20	.15	1.09	.73	.74
	909	.07	.94	30	.14	1.00	.79	.66
Dissimilar	885	.07	.92	10	.15	1.10	.64	.86
	897	.08	.92	20	.15	1.01	.74	.72
	907	.08	.92	30	.14	.98	.78	.66
Partial	804	.13	.87	10	.32	.83	.70	.81
	822	.19	.85	20	.27	.83	.78	.67
	851	.16	.86	30	.22	.84	.81	.62

[a] All anchor item sets contained 20 items

Table 8

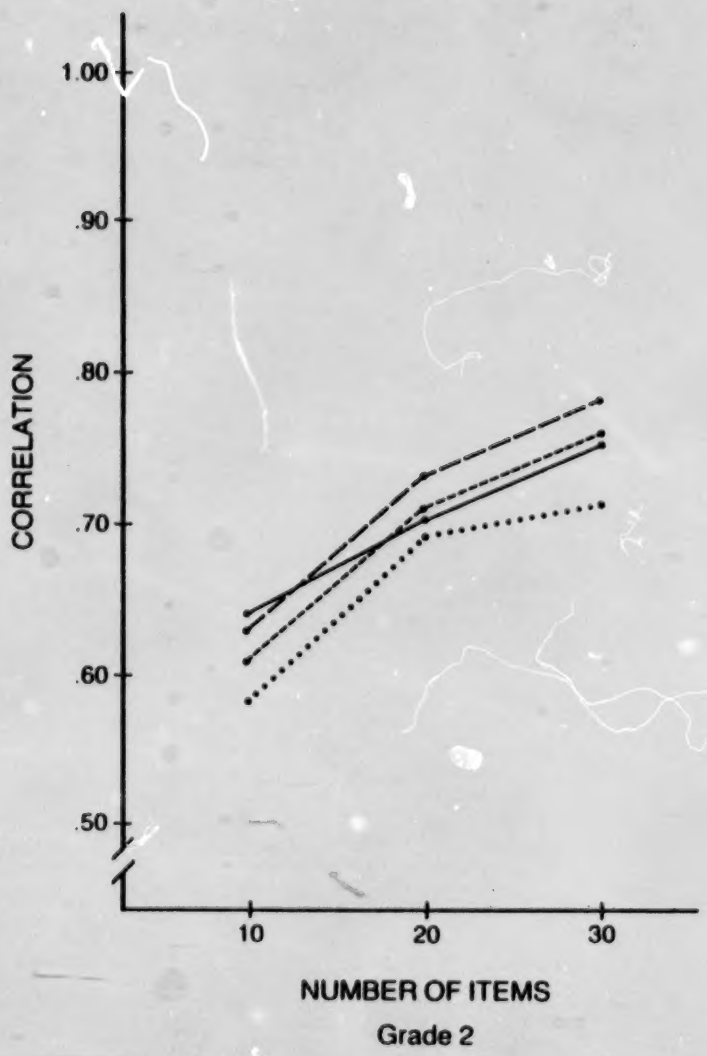
Primary and Indirect Trait Estimate Means, Standard Deviations, Correlations, and Root Mean Squares by Grade (Cross-Validation)[a]

Content-Match Category	N	Primary Trait Estimate $\bar{\hat{\theta}}_1$	Primary Trait Estimate $S_{\hat{\theta}1}$	# Items	Indirect Trait Estimate $\bar{\hat{\theta}}_2$	Indirect Trait Estimate $S_{\hat{\theta}2}$	$r_{\hat{\theta}_1\hat{\theta}_2}$	$\frac{RMSD}{S_{\hat{\theta}}}$
				Grade 2				
Match	544	-.13	.80	10	-.44	.82	.65	.91
	715	.02	.87	20	-.10	1.09	.74	.76
	755	.07	.89	30	-.06	1.09	.77	.71
Similar	544	-.20	.80	10	-.48	.75	.65	.91
	721	-.04	.87	20	-.16	.96	.75	.75
	753	.08	.87	30	-.09	.98	.77	.71
Dissimilar	519	-.20	.81	10	-.60	.91	.53	1.08
	693	.00	.87	20	-.19	1.19	.65	.90
	748	.03	.88	30	-.12	1.21	.69	.84
Partial	644	.00	.87	10	-.18	1.03	.65	.87
	733	-.05	.88	20	-.07	1.15	.70	.81
	774	.08	.89	30	-.04	1.12	.76	.73

Table 8
(continued)

Content-Match Category	N	Primary Trait Estimate $\bar{\hat{\theta}}_1$	Primary Trait Estimate $S_{\hat{\theta}1}$	# Items	Indirect Trait Estimate $\bar{\hat{\theta}}_2$	Indirect Trait Estimate $S_{\hat{\theta}2}$	$r_{\hat{\theta}_1\hat{\theta}_2}$	$\frac{RMSD}{S_{\hat{\theta}}}$
				Grade 6				
Match	835	.04	.83	10	.14	.79	.68	.66
	901	.06	.87	20	.17	.88	.76	.61
	907	.07	.87	30	.20	.87	.80	.56
Similar	903	.05	.89	10	.27	1.10	.65	.88
	927	.04	.90	20	.16	1.08	.74	.74
	928	.04	.90	30	.13	1.02	.79	.65
Dissimilar	894	.06	.88	10	.21	1.02	.61	.87
	917	.05	.89	20	.18	1.02	.76	.68
	917	.06	.89	30	.18	.95	.79	.61
Partial	823	.10	.83	10	.32	.76	.64	.71
	850	.14	.84	20	.27	.80	.75	.59
	873	.12	.85	30	.24	.81	.78	.56

[a] N = 905 for Grade 2 and N = 996 for Grade 6.



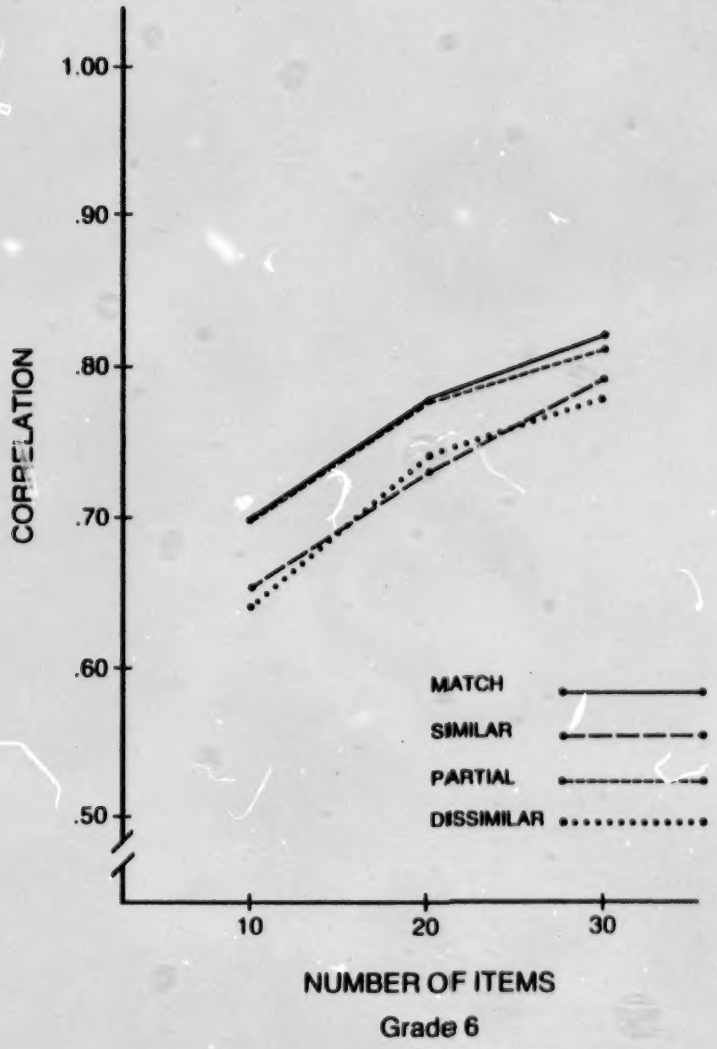


Figure 1. Primary and Indirect Trait Estimate Correlations by Grade (Equating Sample).

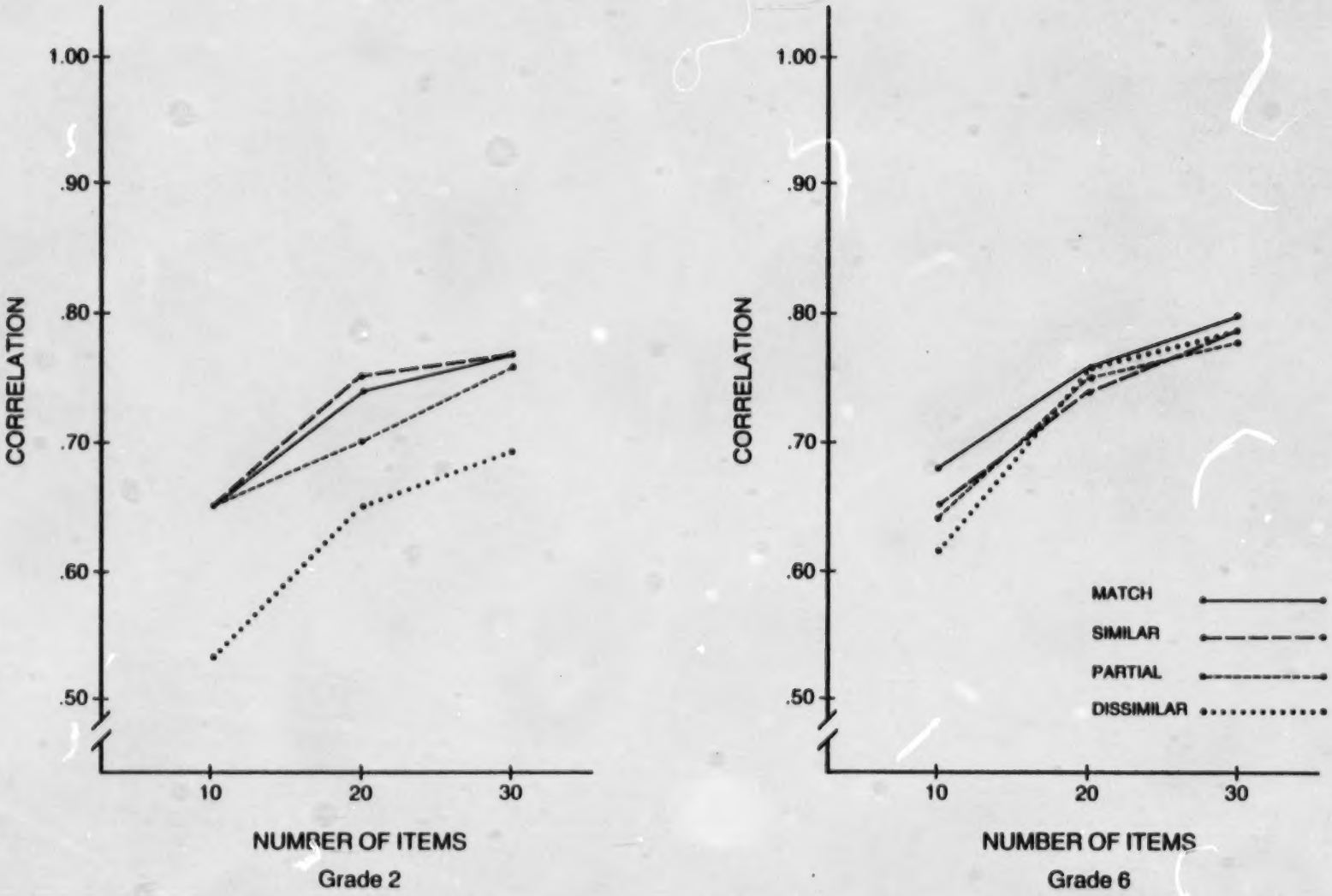


Figure 2. Primary and Indirect Trait Estimate Correlations by Grade (Cross-Validation Sample).